

## **OPAC-Erweiterung durch automatische Indexierung: Eine empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog**

Otto Oberhauser und Josef Labner

### **Abstract**

*In Anlehnung an die in den neunziger Jahren durchgeführten Erschliessungsprojekte MILOS I und MILOS II, die die Eignung eines Verfahrens zur automatischen Indexierung für Bibliothekskataloge zum Thema hatten, wurde eine empirische Untersuchung anhand einer repräsentativen Stichprobe von Titelsätzen aus dem Österreichischen Verbundkatalog durchgeführt. Ziel war die Prüfung und Bewertung der Einsatzmöglichkeit dieses Verfahrens in den Online-Katalogen des Verbundes. Der Realsituation der OPAC-Benutzung gemäss wurde ausschliesslich die Auswirkung auf den mit den automatisch angereicherten Basic Index ("Alle Felder") untersucht. Dazu wurden 100 Suchanfragen zunächst im ursprünglichen Basic Index und sodann im angereicherten Basic Index in einem OPAC unter Aleph 500 durchgeführt. Die Tests erbrachten einen Zuwachs an relevanten Treffern bei nur leichten Verlusten an Precision, eine Reduktion der Nulltreffer-Ergebnisse sowie Aufschlüsse über die Auswirkung einer vorhandenen verbalen Sacherschliessung.*

*In the 1990s the German MILOS projects examined the suitability of an automatic linguistic indexing technique for library OPACs. Following this approach, an empirical investigation was conducted that aimed at assessing and evaluating the use of automatic indexing for the OPACs of the Austrian Library Network. As most users prefer to do their OPAC searches in the basic index the study focused on the effects of enriching this index with automatically generated terms. For this purpose an Aleph 500 OPAC consisting of a representative random sample of records drawn from the Austrian Union Catalogue was used for searching 100 queries in the basic index before and after adding the new index terms. The results include an increase of relevant hits at only moderately lower precision, the reduction of zero-hit results and insights into the role of existing subject headings.*

### **1 Allgemeines, Zielsetzung**

Über die Sinnhaftigkeit der Indexierung (also der inhaltlichen Erschliessung durch das Hinzufügen von Deskriptoren bzw. genormten Schlagwörtern) der Datensätze in Bibliotheks-OPACs besteht, gerade auch im Zeitalter von Google et al., zumindest unter Fachleuten kein Zweifel. In den meisten Bibliotheken des deutschen Sprachraumes geschieht dies bis dato ausschliesslich manuell bzw. intellektuell – im wissenschaftlichen Bereich seit rund 15 Jahren überwiegend durch Anwendung der *Regeln für den Schlagwortkatalog* (RSWK) bzw. der *Schlagwortnormdatei* (SWD).<sup>1</sup> Da die intellektuelle Indexierung trotz Fremddatennutzung und kooperativer Katalogisierung zeit- und personalintensiv ist, wird sie von vielen Bibliotheken nur für einen Teil ihrer Bestände bzw. Zugänge durchgeführt, wodurch der oft nicht unerhebliche Rest bestenfalls über Stichwörter aus Titel/Untertitel sachlich

recherchierbar ist (und meist auf Dauer so verbleibt).<sup>2</sup> In vielen OPACs führt daher die Recherche im Schlagwortindex zu einer für die Benutzer undurchschaubaren Limitierung der Suche auf ein undefiniertes Sub-Set von Einträgen, sodass viele Bibliotheken dazu übergegangen sind, nicht den Schlagwortindex, sondern den "Basic Index" (der alle inhaltlich relevanten oder gar überhaupt alle Datenfelder, in der Regel in verstichworteter Form, enthält) als prominentesten Sucheinstieg ihres OPACs anzubieten.<sup>3</sup>

Erst in wenigen deutschsprachigen Bibliotheken werden auch Verfahren der *automatischen* bzw. *maschinellen* Indexierung eingesetzt. Darunter sind all jene Verfahren zu verstehen, "die vollautomatisch Dokumente analysieren und abgeleitet aus dieser Analyse entweder ausgewählte Terme aus dem Dokument extrahieren und – unter bestimmten Verfahrensvoraussetzungen in einer bearbeiteten Form – als Indexterme abspeichern (Extraktionsverfahren) oder Deskriptoren einer kontrollierten Indexierungssprache dem Dokument als Inhaltsrepräsentanten zuweisen (Additionsverfahren)".<sup>4</sup> Für deutschsprachige OPACs – also solche, in denen die Mehrzahl der *Katalogisate* (nicht notwendigerweise der Werke) in deutscher Sprache vorliegt und bei denen sich das inhaltstragende Wortmaterial in der Regel auf Titel- bzw. Schlagwörter beschränkt – eignen sich primär *linguistische* Indexierungsverfahren.<sup>5</sup> Diese gehen von der Hypothese aus, dass die meisten (verstichworteten) Terme zu heterogen seien, um "gute" Suchbegriffe darzustellen, und versuchen daher, mit verschiedenen Methoden – Reduktion auf Grundformen, Zerlegung von Komposita usw. – eine Standardisierung bzw. Normierung dieses Vokabulars zu erzielen. Während dies etwa im Fall des Englischen, also einer Sprache mit stark regelhafter Wortbildung, mit relativ einfachen Techniken erreicht werden kann, benötigt man für deutschsprachige Texte, die durch Mehrzahlformen (Haus – Häuser), Flexionsformen (Häusern), Kompositabildungen (Krankenhäusern) und weitere Unregelmässigkeiten gekennzeichnet sind, Verfahren, die auf sehr umfangreichen *Wörterbüchern* beruhen, in denen die konkret auftretenden Wortformen und ihre gewünschte Umsetzung mehr oder weniger detailliert lexikalisiert sind. Dabei genügt es, *wortbezogen* vorzugehen, da angesichts des geringen Umfangs des Wortmaterials der Titelaufnahmen string- und satzorientierte Verfahren nicht unbedingt benötigt werden.

"Pionier" der automatischen Indexierung deutschsprachiger OPACs ist KLAUS LEPSKY, der in den 1990er Jahren an der ULB Düsseldorf die beiden MILOS-Projekte durchführte,<sup>6</sup> im Rahmen derer die Eignung des wörterbuchbasierten Verfahrens IDX<sup>7</sup> für deutsche OPACs getestet wurde. Während im Projekt MILOS I (1993–1995) 40.000 Titel aus dem OPAC der ULB Düsseldorf maschinell indexiert und einem Retrievaltest mit 50 Suchfragen unterzogen wurden,<sup>8</sup> basierte das MILOS II-Projekt (1995–1996) auf der Indexierung von rund 190.000

Titelsätzen der Deutschen Nationalbibliographie, der Einbindung der SWD in das Indexierungsverfahren sowie auf einem Retrievaltest mit 100 Suchfragen in einer Allegro-Umgebung.<sup>9</sup> Ein weiterer Test mit ca. 47.000 ekz-Datensätzen und 30 Suchfragen wurde im Rahmen einer bibliothekarischen Diplomarbeit in Bonn durchgeführt.<sup>10</sup> Alle drei Studien gelangten – verallgemeinert formuliert – zu dem Schluss, dass durch die automatische Indexierung eine beträchtliche Steigerung des Recalls bzw. der Menge relevanter Treffer erzielt wurde, ohne dass es gleichzeitig zu dramatischen Einbussen hinsichtlich der Precision kam. Ausserdem hatte die automatische Indexierung zur Folge, dass die Zahl der Nulltreffer-Resultate, die ja das Problem vieler OPAC-Benutzer darstellen, deutlich zurückging.

Trotz dieser Erfolge blieb, wie oben angedeutet, der praktische Einsatz der automatischen Indexierung bislang auf wenige Bibliotheken beschränkt. Über die Gründe für diese Zurückhaltung mag man spekulieren;<sup>11</sup> Fakt ist, dass zur Zeit nur die ULB Düsseldorf (allerdings noch nicht im aktuellen Aleph-OPAC), die Bibliothek des Zentralinstituts für Kunstgeschichte, München sowie die Bibliothek der Friedrich-Ebert-Stiftung, Bonn eine automatische Indexierung mit MILOS/IDX im Routinebetrieb durchführen.<sup>12</sup> Die Deutsche Bibliothek plant, dieses Verfahren ab 2004 zur Erschliessung ihres Retro-Kataloges (1945 ff.) einzusetzen. Als erste österreichische Einrichtung begann kürzlich die Vorarlberger Landesbibliothek, Bregenz, quasi in der Nachfolge des ebenfalls in Düsseldorf (1997–1998) durchgeführten "Kascade-Projekts",<sup>13</sup> mit der maschinellen Indexierung von gescannten Inhaltsverzeichnissen von Monographien und Sammelwerken, wobei das Programm CAI-Engine zum Einsatz gelangt.<sup>14</sup>

Der Österreichische Bibliothekenverbund, an dem derzeit 58 wissenschaftliche, administrative und Spezialbibliotheken teilnehmen,<sup>15</sup> verfügt über einen OPAC mit 3,76 Millionen Titel- und 6,78 Millionen Exemplardatensätzen, der seit 1999 unter der Software "Aleph 500" verwaltet wird.<sup>16</sup> Die verbale Sacherschliessung erfolgt auf der Basis von RSWK und SWD; allerdings weisen nur 42,5% der Datensätze eine solche Erschliessung auf. Da das System Aleph 500 es ermöglicht, zu den Schlagwörtern der in den MAB-Kategorien 902, 907, ..., 947 enthaltenen RSWK-Ketten auch die Verweisungsformen aus der SWD einzuspielen, weisen die entsprechenden Suchregister (Schlagwortindex bzw. "Alle Felder" = Basic Index) auch diese Begriffe auf,<sup>17</sup> was durchaus als *rudimentäre* Form einer automatischen Indexierung bezeichnet werden könnte.

Das Interesse an einer *weiterreichenden* automatischen Indexierung liegt angesichts der nur partiellen intellektuellen sachlichen Erschliessung der Datensätze nahe. Warum, so könnte

man fragen, sollte jedoch angesichts der ermutigenden Resultate der deutschen Studien noch ein weiterer Test zur Prüfung und Bewertung der zukünftigen Einsatzmöglichkeit eines linguistischen Verfahrens zur Anreicherung des Suchvokabulars mit zusätzlichen Suchbegriffen ("Indexaten") vonnöten sein? In der Zentrale des Österreichischen Bibliothekenverbundes waren wir an mehreren Aspekten interessiert:

- Einsatz des Verfahrens in einem realen OPAC
- Einsatz in einer Aleph 500-Umgebung
- Einsatz in einer "Alle Felder"-Suche durch Vergleich der Retrievalergebnisse *vor* und *nach* der Anreicherung des Basic Index durch maschinell gewonnene Indexate<sup>18</sup>

Die in den MILOS-Studien gewonnenen Hypothesen, wonach die automatische Indexierung eine erhöhte Ausbeute relevanter Treffer bei nur wenig reduzierter Precision sowie eine Reduktion der trefferlosen Suchergebnisse bewirken würde, sollten der Studie natürlich ebenfalls zugrundegelegt und abermals geprüft werden.

## **2 Methode & Testumgebung**

### **2.1 Verfahren (MILOS/IDX)**

Im Herbst 2002 hatte einer der Autoren Gelegenheit, an der FH Köln, wo K. LEPSKY seit 2001 als Professor wirkt, an einem Laborpraktikum zum Thema "automatische Indexierung" teilzunehmen. Durch diesen persönlichen Kontakt ergab sich die Möglichkeit, Daten aus dem Österreichischen Verbund-OPAC einer automatischen Indexierung mit dem wörterbuchbasierten linguistischen Verfahren MILOS/IDX zu unterziehen. Der Leistungsumfang von MILOS/IDX kann kurz wie folgt beschrieben werden:<sup>19</sup>

- Grundformenreduktion (Lemmatisierung), z.B. Häuser → Haus; gegangen → gehen
- Zerlegung von Komposita (Dekomposition), z.B. Haustür → Haus, Tür; Baumangel → Bau, Mangel (nicht: Baum, Angel); Staatsexamen → Staat, Examen
- Wortableitung (Derivation), z.B. französisch → Frankreich; Besichtigung → besichtigen
- Mehrworterkennung (Phrasierung), z.B. sozialer Wandel; österreichische Literaturzeitschrift; Kolloquium über Information (diese Funktion war für den vorliegenden Test nicht von Bedeutung, da solche Gruppen beim Aufbau des Basic Index in einzelne Stichwörter zerlegt werden)

- Wortbindestrichergänzung, z.B. Geistes- und Kulturgeschichte → Geistesgeschichte; Brücken- und Ingenieurbau → Brückenbau
- Zuordnung diskontinuierlicher Verbteile, z.B. kamen ... an → ankommen
- Markierung und Eliminierung von Stoppwörtern (für den vorliegenden Test nicht relevant)
- Einbindung der SWD bzw. eines Thesaurus für Wortrelationierungen (im Gegensatz zur bisher üblichen Vorgangsweise im Österreichischen Verbundkatalog auch für Stichwörter aus dem Titel), z.B. Schlepper → Ackerschlepper, Traktor, Trecker, Radschlepper, Reifenschlepper, Radtraktor (Verweisungsformen der SWD)
- Wortbezogene Übersetzung bzw. sprachspezifische Verarbeitung auch in Englisch und Französisch (wurde in den MILOS-Studien und im vorliegenden Test nicht verwendet)

## 2.2 Stichprobe / Testdaten

Der Retrievaltest sollte anhand einer für den Österreichischen Verbundkatalog repräsentativen Stichprobe von Titeldatensätzen durchgeführt werden, deren Grösse mit ca. 100.000 festgesetzt wurde (vgl. Tab. 1). Zu diesem Zweck wurden mit Hilfe eines Zufallszahlengenerators 110.000 Systemnummern (= eindeutige Identifikatoren der bibliographischen Datensätze in Aleph 500) generiert, von denen nach Ausscheidung der a priori schwer abschätzbaren Menge doppelt erzeugter Zufallszahlen sowie der "Löschsätze" (= im System verbleibende kategorienlose Torsi gelöschter Datensätze) eine Stichprobe von insgesamt 106.904 Datensätzen verblieb. Von diesen wies der grösste Teil (97.460) zumindest eine der MAB-Kategorien 331 (Hauptsachtitel), 335 (Zusätze zum HST) oder 902, 907, ...947 (RSWK-Schlagwortketten) auf und verfügten damit über Wortmaterial, das als Basis für die automatische Indexierung dienen konnte.

Grundgesamtheit (Verbundkatalog, März 2003):	ca. 3,6 Mio. Datensätze
Angestrebter Stichprobenumfang:	ca. 100.000
Erzeugte Zufallszahlen (Systemnummern):	110.000
davon doppelt u. ausgeschieden:	1.667
Bruttoansatz Stichprobe:	108.333
ohne Löschsätze:	106.904
zu indexierende MAB-Kategorien vorhanden	97.460

*Tab. 1: Erstellung der Stichprobe*

Die Stichprobe wurde aus dem Verbundkatalog exportiert und in einen neuen Aleph-OPAC geladen. Sie ist, wie eine statistische Auswertung zeigte, hinsichtlich zahlreicher Parameter ein nahezu exaktes Abbild des Verbundkataloges (Tab. 2).

	Verbundkatalog	Stichprobe
Anzahl absolut	3,594.154	106.904
Hauptsachtitel (HST) vorhanden	88,6 %	88,7 %
Zusätze zum HST vorhanden	42,6 %	42,5 %
Beschlagwortet (RSWK)	42,5 %	42,6 %
Sprache Deutsch	60,8 %	61,2 %
Sprache Englisch	21,4 %	21,3 %
Erscheinungsland USA	10,5 %	10,4 %
Erscheinungsjahr 1990	2,9 %	2,9 %
Erscheinungsjahr 2000	3,2 %	3,2 %
Monographien	39,3 %	39,1 %
Zeitschriften & Serien	9,8 %	9,9 %
MAB-Hauptsätze	81,8 %	82,0 %
Katalogisiert durch UB Wien	14,2 %	14,0 %

*Tab. 2: Vergleich der Stichprobe mit dem Verbundkatalog hinsichtlich ausgewählter Parameter*

Die erwähnten 97.460 Datensätze wurden in einem auf die relevanten Kategorien reduzierten Exportformat an die FH Köln übermittelt und dort in einer Allegro-Umgebung der Indexierung mit MILOS/IDX zugeführt. Nach einer maschinellen Spracherkennung – dabei geht es um die deutsche Sprache der Katalogisate, nicht der Dokumente – und der Ausscheidung von nicht indexierbaren Datensätzen (Bandaufführungen, Bindeeinheiten, Titel ausschliesslich Eigenname / Geographikum usw.) qualifizierten sich insgesamt 72.006 Datensätze für die automatische Indexierung.

## 2.3 Beispielsätze

Nach der Durchführung der automatischen Indexierung (Mai 2003) wurden die um zwei neue Kategorien – Indexate auf Basis von Titel/Untertitel; Indexate auf Basis der Schlagwörter – angereicherten Datensätze an die Wiener Verbundzentrale übermittelt und dem unter Aleph 500 erstellten Test-OPAC hinzugefügt. Im Beispieldatensatz in Abb. 1 wird veranschaulicht, welche Indexate auf Basis welcher Kategorien erstellt wurden und welche Funktionalität des Verfahrens (z.B. Grundformenreduktion, Wortableitung) dabei jeweils zum Tragen kam. In diesem Beispiel fällt etwa auch positiv auf, dass bei der automatischen Indexierung die Einbindung der SWD nicht soweit ging, auch Geographika zu relationieren, wodurch hier die Hinzufügung der historischen Synonyme für Österreich (Ostmark <1938–1945>, Cisleithanien, Habsburgerreich) unterblieb.

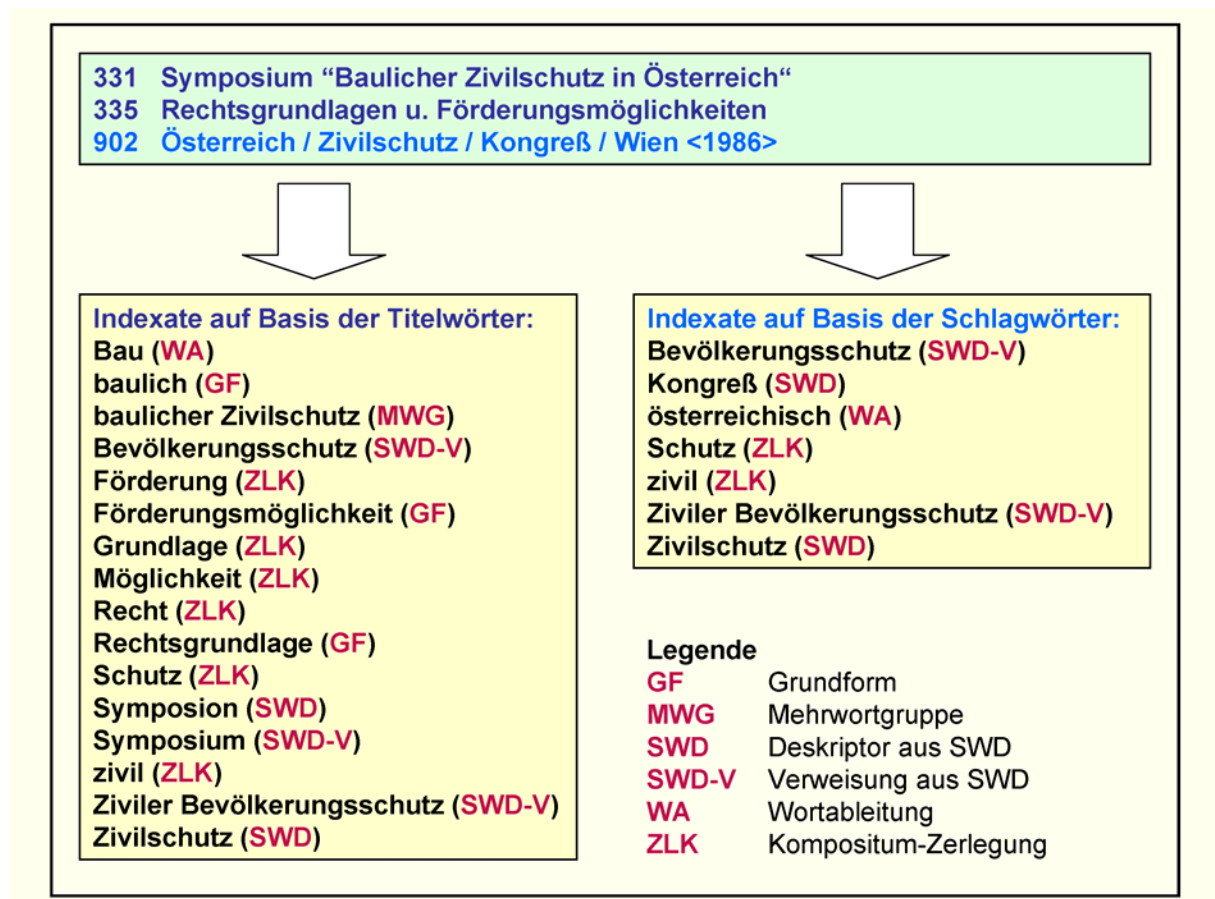


Abb. 1: Beispiel für einen automatisch indexierten Testdatensatz

Dass die automatische Indexierung in vielerlei Hinsicht an die derzeitigen Beschränkungen der EDV gebunden ist bzw. "dort ihre Grenzen erreicht, wo die Intelligenz beginnt" (LEPSKY), wird aus den in Abb. 2 dargestellten Fehlermöglichkeiten ersichtlich. Im oberen Beispiel kommt es aufgrund des Eigennamens "Hahn" zur falschen Erkennung eines SWD-Deskriptors ("Hahn") bzw. dadurch zur Hinzufügung einer störenden SWD-Verweisungsform ("Haushahn"); im unteren Beispiel passiert Ähnliches (SWD-Deskriptor "Reihe"; SWD-Verweisungsformen "Progression <Mathematik>" und "Unendliche Reihe"). Beide Beispiele zeigen die Problematik des Fehlens einer semantischen Disambiguierung; allerdings erwies sich der Anteil solcher und ähnlicher Fehlindexierungen in den Testdaten als eher gering.

## 2.4 Test-OPAC

Beim Einspielen der neuen Indexate in den Testdatenbestand wurde neben dem bestehenden Basic Index ("Alle Felder") ein zweiter Basic Index ("Alle Felder inkl. neue Indexate") aufgebaut,<sup>20</sup> wobei alle Mehrwortgruppen in Einzelstichwörter umgewandelt wurden. Dies betraf bedauerlicherweise auch die SWD-Homonyme bzw. deren Zusätze, was allerdings auch für den bereits bestehenden Basic Index des Verbundkataloges zutrifft.



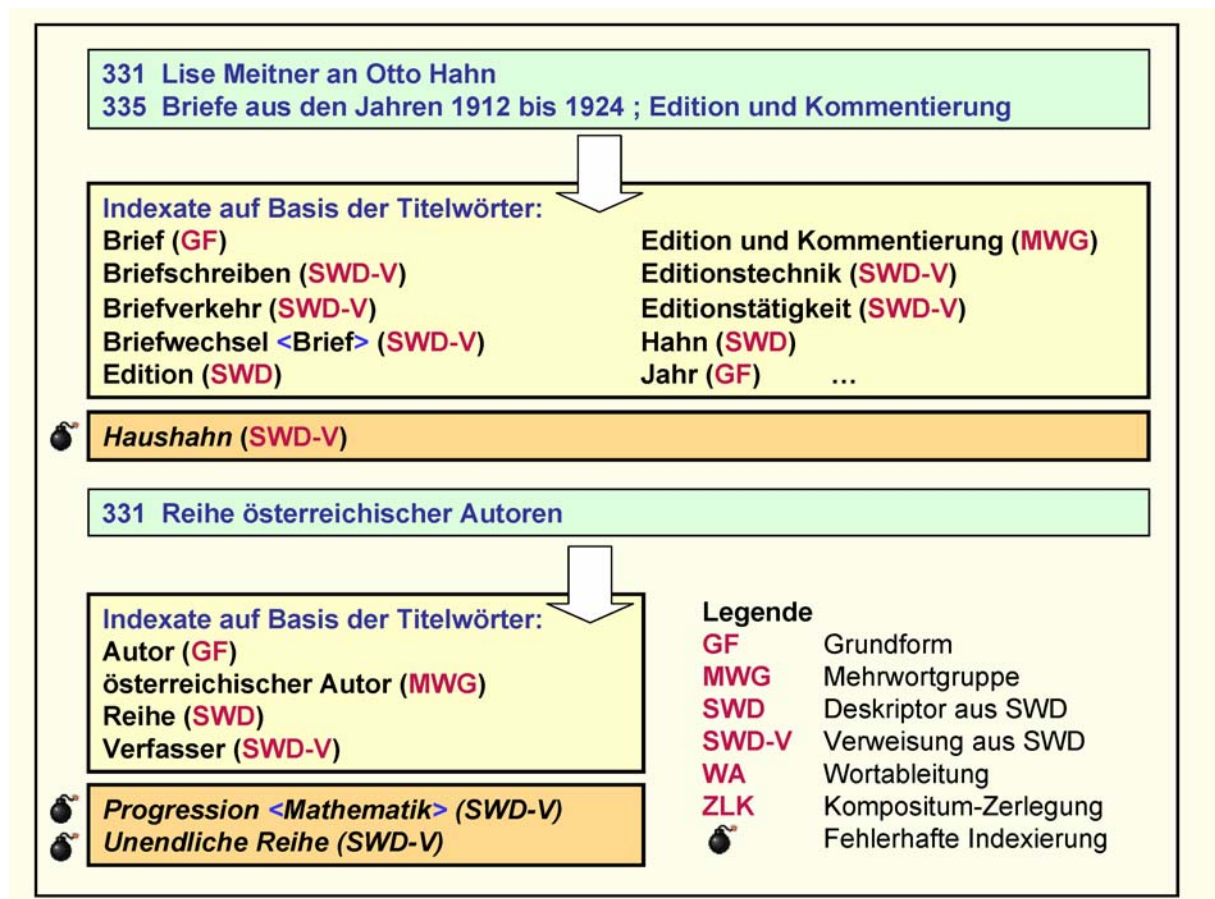


Abb. 2: Fehlermöglichkeiten bei der automatischen Indexierung

Abb. 3 zeigt die Suchmaske des für den Retrievaltest eingerichteten OPACs. Im Gegensatz zur Standardoberfläche des Österreichischen Verbund-OPACs verfügte dieser Katalog über zwei "Alle Felder"-Suchmöglichkeiten:

- Schnellsuche in allen Feldern (wie bisher)
- Schnellsuche in allen Feldern inkl. neue Indexate

Die sonstigen OPAC-Funktionalitäten wurden auf die für den Test nötigen Features reduziert, sodass z.B. Indexblättern, Mehrfeldersuche und Expertensuche wegfielen. Um den Testpersonen die Navigation soweit wie möglich zu erleichtern, wurde das Anzeigelimit für die "Ergebnisliste" (Kurztitelanzeige) von üblicherweise 10 auf das Maximum von 99 erhöht, was zwar die Performance dieses Aleph-OPACs deutlich beeinträchtigte, aber den Testpersonen das Weiterblättern bei grösseren Treffermengen weitgehend ersparte.



**OPAC für Retrievaltest Automatische Indexierung**

[Neustart](#) | [Katalogauswahl](#) | [Optionen](#) |

[Suchen](#) | [Ergebnisliste](#) | [Suchgeschichte](#) | [Bücherkorb](#)

**Schnellsuche in allen Feldern**

**Schnellsuche in allen Feldern mit Indexaten**

*Abb. 3: Suchmaske des Test-OPACs*

### **3 Durchführung der Retrievaltests**

#### **3.1 Suchfragen**

Nach der Anreicherung des Test-OPACs mit den neuen Indexaten konnte mit der praktischen Durchführung der Retrievaltests begonnen werden. Für diese Tests wurden die 100 Suchfragen aus MILOS II herangezogen<sup>21</sup>; dies geschah vor allem aus arbeitsökonomischen Gründen. Die 100 Fragen basierten zur Hälfte auf realen Benutzerrecherchen der ULB Düsseldorf und hatten bereits dem Retrievaltest in MILOS I als Queries gedient, während die anderen 50 Fragen im Rahmen von MILOS II zur Abdeckung diverser Typen von Benutzerrecherchen in OPACs von Universalbibliotheken entwickelt wurden. Die einzige Ausnahme bildete der Suchbegriff "Internet", der im Gegensatz zu den 1990er Jahren heute eine zu grosse Treffermenge erbringen würde; als Ersatz wurde für die vorliegende Studie der Begriff "Intranet" verwendet.

Während sich die MILOS-Studien mit der Leistungsfähigkeit einzelner (Allegro-)Register befassten, konzentrierte sich unser Projekt ausschliesslich auf die Suche im Basic Index ("Alle Felder"). Daher war es erforderlich, die 100 Suchfragen für die Recherche in diesem Index aufzubereiten. Bei einer solchen Suche ist es im Gegensatz zur

"registeraufblättern" Allegro-Recherche nicht möglich, verschiedene vom System angebotene Suchformulierungen zu verwenden und dann die für das jeweilige Register beste für die Auswertung heranzuziehen;<sup>22</sup> vielmehr muss praktisch "blind" im verstichworteten Vokabular des Basic Index gesucht werden. Da nicht bekannt ist, welche Arten von Suchformulierungen von Benutzern verwendet werden, wurde entschieden, auf jeglichen Versuch der Simulation eines konkreten Benutzerverhaltens zu verzichten und statt dessen nur *einfachste* mögliche Eingabevarianten zu berücksichtigen. So wurden weder trunkierte Suchbegriffe noch Oder-Verknüpfungen von (Quasi-)Synonymen bzw. Unterbegriffen verwendet, obwohl aus Logfileanalysen bekannt ist, dass derlei in der Praxis durchaus vorkommt. Daher wurden in jenen Fällen, wo dies sinnvoll erschien, Suchformulierungen mit Oder-Verknüpfungen von Einzahl- und Mehrzahlformen vorgesehen (dies betraf jedoch nur weniger als die Hälfte der Queries). Tab. 3 zeigt einige Beispiele für Suchfragen und ihre Aufbereitung für die Suche im Basic Index des neu eingerichteten Aleph-Test-OPACs.

Fragestellung	Aufbereitung für Test-OPAC
Hemmung von Enzymen	hemmung AND (enzym OR enzyme)
Umweltökonomie/-ökonomik	umweltökonomie OR umweltökonomik
Ergonomie	ergonomie
Alternative Energiequellen	alternative AND (energiequelle OR energiequellen)
Zimmerpflanzen	zimmerpflanze OR zimmerpflanzen
Strassenkinder	strassenkind OR strassenkinder
Selbstbewusstsein stärken	selbstbewusstsein AND (stärken OR stärkung)
Denken und Lernen	denken AND lernen
Medizin im Dritten Reich	medizin AND drittes AND reich

Tab. 3: Beispiele für Fragestellungen aus MILOS II und ihre Aufbereitung für die "Alle Felder-Suche" im Aleph-OPAC

Für die Bewertung von Retrievalergebnissen werden üblicherweise die Parameter Recall und Precision herangezogen. Da die Berechnung des ersten Parameters die Kenntnis der im Datenbestand vorhandenen relevanten Dokumente voraussetzt (was bei grossen Testdatenbeständen kaum möglich ist), wurde – wie im Retrievaltest MILOS II – auf die Berechnung des Recalls verzichtet und nur die Zahl bzw. der Zuwachs relevanter Treffer betrachtet. Hinsichtlich der Bestimmung der Precision wurde das "weitgefasste" Relevanzkriterium aus MILOS II herangezogen – jeder Treffer, bei dem ein Interesse am Originaldokument vermutet werden kann, sollte als "relevant" gelten.<sup>23</sup>

### 3.2 Retrievaltests

Die Studie umfasste insgesamt drei Retrievaltests:

- Der erste Test wurde im Juni 2003 durchgeführt. Da der dafür erforderliche Arbeitsaufwand mangels einschlägiger Erfahrungen überschätzt worden war, erfolgte eine Aufteilung der Recherchen auf die Autoren und zehn weitere Versuchspersonen (vor allem aus der Verbundzentrale); die Tests benötigten jedoch einen geringeren Zeitaufwand als ursprünglich erwartet. Jede Testperson erhielt 8 bis 9 vorformulierte Suchfragen zur Eingabe im Basic Index des Test-OPACs *ohne* bzw. *mit* Indexaten, wobei die jeweilige Trefferzahl sowie die Zahl der nach dem obigen Kriterium identifizierten relevanten Treffer zu protokollieren waren.
- Ein zweiter Retrievaltest, bei dem es um die Differenzierung zwischen beschlagworteten und nicht beschlagworteten Datensätzen ging, wurde im Juli 2003 durch die Autoren durchgeführt. Diese Differenzierung hätte eigentlich schon im ersten Test getroffen werden können, doch konkretisierte sich das Interesse an der dahinterliegenden Fragestellung (Rolle der Beschlagwortung beim Einsatz der automatischen Indexierung) erst nach Vorliegen der Ergebnisse des ersten Tests.
- Im Oktober 2003 wurde schliesslich ein dritter Retrievaltest (wieder durch die Autoren) durchgeführt. Dabei ging es um die Überprüfung der Hypothese, wonach die in den ersten beiden Tests verwendeten Suchformulierungen die Effekte der automatischen Indexierung reduziert hätten, da neben Singular- auch Pluralformen und (vereinzelt) Wortableitungen verwendet worden waren. Aus diesem Grunde wurden alle in Frage kommenden Suchformulierungen identifiziert (43 von 100) und für diese Fälle erneute Abfragen unter Verwendung nur der Singularform bzw. nur der Pluralform (letzteres lediglich in einzelnen begründeten Fällen wie "Kurden", "Zeugen Jehovas") durchgeführt.

## **4 Testergebnisse**

### **4.1 Retrievaltest 1**

Das Ergebnis des ersten Retrievaltests ist in Abb. 4 (Verteilung der Trefferzahlen für die Suchen ohne bzw. mit Indexaten in kategorisierter Form) und Abb. 5 (durchschnittliche Trefferzahlen und Relevanz) dargestellt. Die Suche im herkömmlichen Basic Index erbrachte durchschnittlich 10,65 Treffer je Suchfrage, wobei eine bemerkenswert hohe Precision erreicht wurde (9,78 von 10,65 = 91,8%). In dem mit den neuen Indexaten angereicherten Basic Index wurden durchschnittlich 14,10 Dokumente gefunden; der Zuwachs von somit durchschnittlich 3,45 Titeln entspricht einem Anstieg um 32,4%. Durchschnittlich 1,94 neue Treffer pro Suchfrage erwiesen sich als relevant, was – bezogen auf das Gesamtergebnis –

eine nur mässige Abnahme der Precision bedeutet (von ursprünglich 91,8% auf 83,1%). Die Zahl der Nulltreffer-Resultate konnte um etwa ein Drittel vermindert werden (von 16 auf 11).

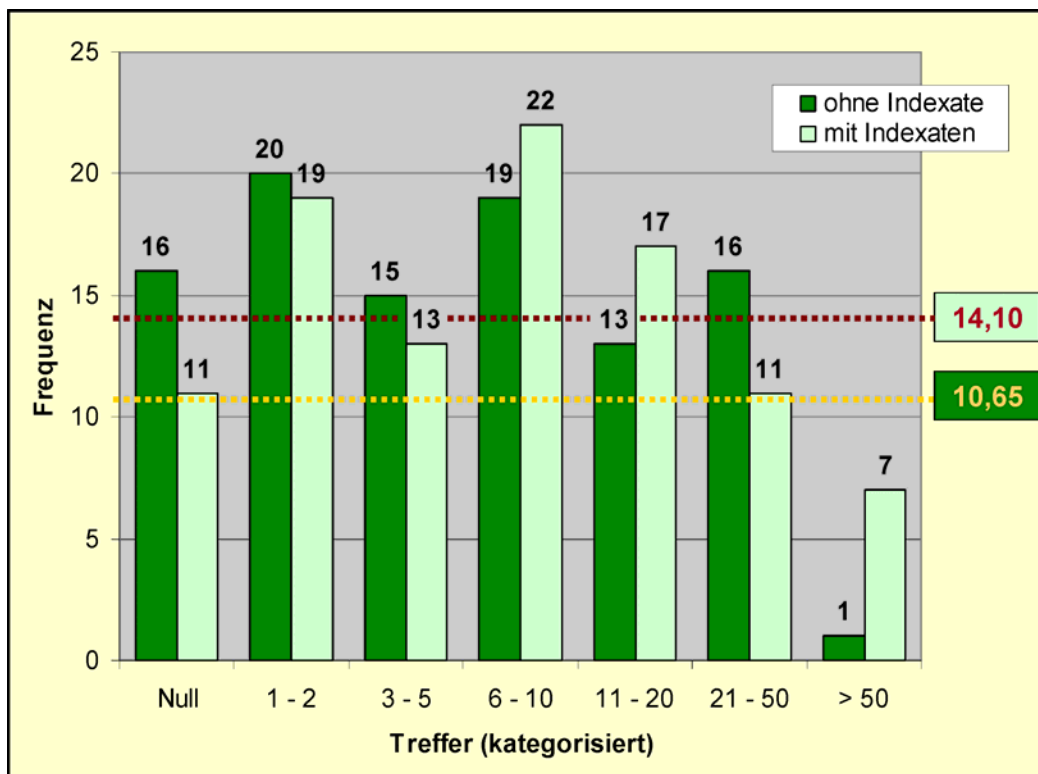


Abb. 4: Verteilung der Trefferzahlen

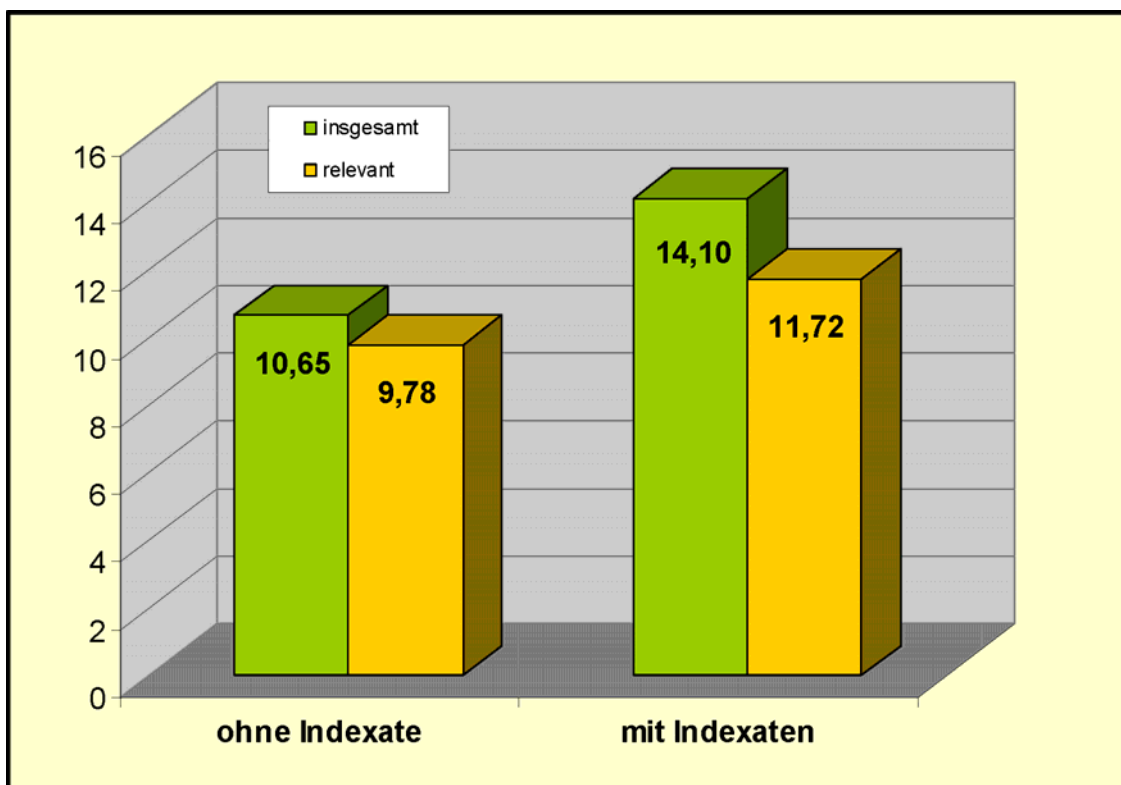


Abb. 5: Durchschnittliche Trefferzahlen und Relevanz

## 4.2 Retrievaltest 2

Im zweiten Retrievaltest wurden die gefundenen Dokumente nach beschlagworteten bzw. nicht beschlagworteten differenziert. Wie bereits erwähnt, betrug der Anteil der Datensätze mit RSWK-Ketten in der Stichprobe insgesamt 42,6%. Nun stellte sich heraus, dass dieser Anteil unter den Treffern ohne Indexate mehr als doppelt so hoch war (90,7% = 966 von 1.065 Treffern, vgl. Abb. 6). Mit anderen Worten, die verwendeten Suchbegriffe führten zunächst vor allem zu beschlagworteten Titeln. Sofern nicht beschlagwortete Titel überhaupt gefunden wurden (99 von 1.065), waren diese ebenfalls von hoher Relevanz.

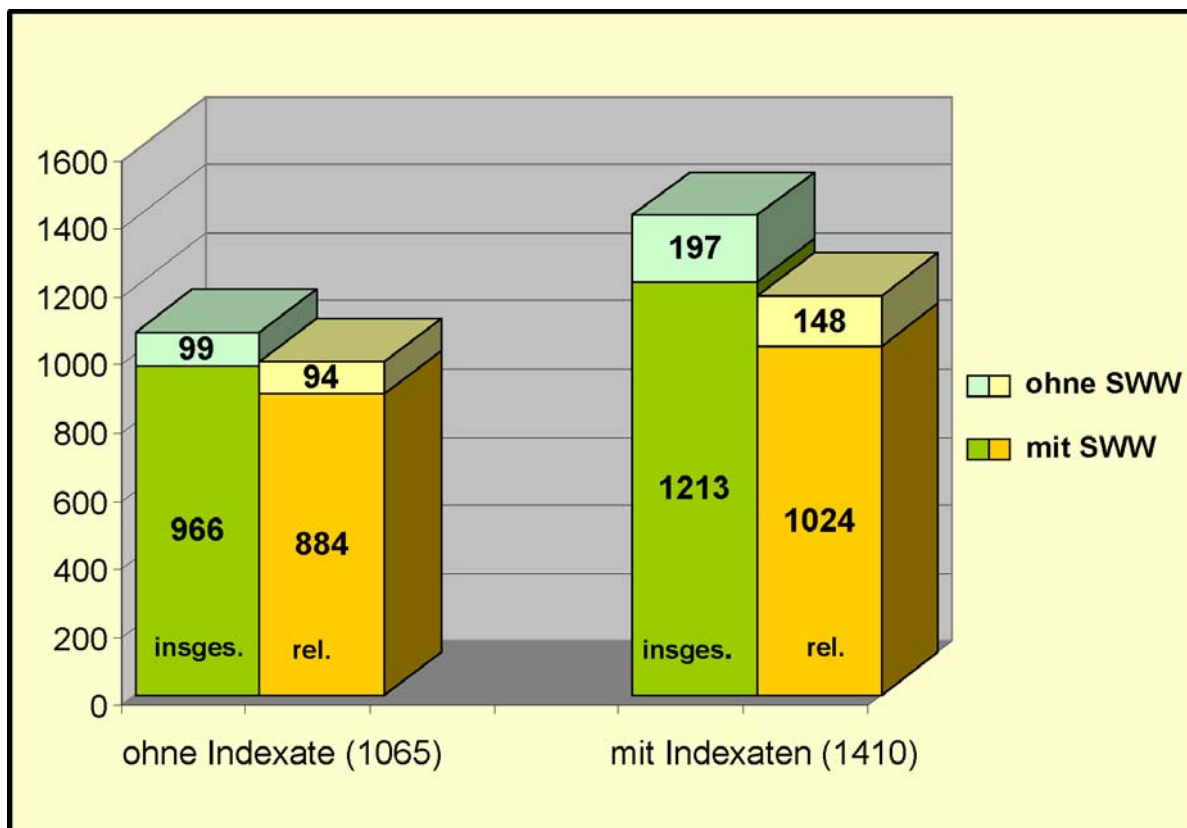


Abb. 6: Trefferzahlen nach beschlagworteten / nicht beschlagworteten Datensätzen und Relevanz

Nach der automatischen Indexierung wurden die verwendeten Suchbegriffe vermehrt auch bei nicht beschlagworteten Titeln gefunden (nunmehr 197; zuvor 99), was die Effekte der automatischen Indexierung bei den Stichwörtern aus dem Titel illustriert (z.B. Grundformen, zerlegte Komposita, Wortableitungen, Synonyme aus der SWD).

	Zuwachs absolut	Prozentwerte
Zuwachs insgesamt	345	32,4 % von 1.065 (Treffer ohne Indexate)
davon relevant	194	56,2 % von 345
Beschlagwortet (RSWK)	247	71,6 % von 345
davon relevant	140	56,7 % von 247
Nicht beschlagwortet	98	28,4 % von 345
davon relevant	54	55,1 % von 98

Tab. 4: Absoluter Zuwachs (Aufteilung)

Tab. 4 zeigt eine Analyse der 345 Datensätze, die sich bei der zweiten Suche, d.h. auf Basis der neuen Indexate, zusätzlich als Treffer qualifizierten. Der Anteil der beschlagworteten Treffer ist wiederum erstaunlich hoch (274 = 71,6 % der 345 hinzugekommenen Titel). Schon vor der automatischen Indexierung waren ja im Basic Index die Synonyme aus der SWD vorhanden gewesen, sodass wir befürchtet hatten, dass das Verfahren im Österreichischen Verbundkatalog keinen nennenswerten *beschlagworteten* Titelzuwachs mehr erbringen würde. Auch unter diesen Treffern ist allerdings – naheliegenderweise – ein spürbarer Rückgang der Precision zu verzeichnen (56,7% im Vergleich zu ursprünglich 91,8%). Der Anteil der relevanten Treffer unter den neu hinzugekommenen *nicht beschlagworteten* 98 Titeln (28,4 %) ist mit 55,1% dagegen höher als erwartet.

Die beiden Beispiele in Tab. 5 dienen der Veranschaulichung der Retrievaltests 1 und 2. Im ersten Beispiel (Frau und Beruf) verdoppelte sich die Zahl der Treffer im Basic Index mit neuen Indexaten, wobei die Precision bemerkenswerterweise anstieg. Im zweiten Beispiel (Geschichte des Mittelmeerraums) kam es sogar zu einer Verfünffachung der Treffermenge, allerdings bei einem spürbaren Rückgang des Anteils relevanter Treffer. Im Hinblick auf die Beschlagwortung spiegeln beide Beispiele in etwa das Gesamtergebnis wider, d.h. dass zunächst fast nur beschlagwortete Titel gefunden wurden.

Beispiel 1: Frau und Beruf				
(frau OR frauen) AND beruf	ohne Indexate		mit Indexaten	
	SWW	oSWW	SWW	oSWW
Treffer	23		48	
	21	2	44	4
Relevant	17	1	40	3
	18 (78,3%)		43 (89,6%)	

Beispiel 2: Geschichte des Mittelmeerraums				
geschichte AND mittelmeerraum	ohne Indexate		mit Indexaten	
	SWW	oSWW	SWW	oSWW
Treffer	30		151	
	30	0	143	8
Relevant	22	0	70	4
	22 (73,3%)		74 (49,0%)	

Legende: SWW = mit RSWK-Ketten; oSWW = ohne RSWK-Ketten				
---	--	--	--	--

Tab. 5: Veranschaulichung der Retrievaltests durch ausgesuchte Beispiele

Dass die "rudimentäre automatische Indexierung", wie sie im Österreichischen Verbundkatalog durch Hinzunahme der Synonyme aus der SWD bereits durchgeführt wird, durch ein Verfahren wie MILOS/IDX durchaus noch verbessert werden kann, zeigt Abb. 7 anhand eines Datensatz aus Beispiel 1 (Tab. 5). Der für den Sucherfolg entscheidende Indexterm "Beruf" entstand hier erst nach Zerlegung des Kompositums "Berufstätigkeit" durch das automatische Indexierungsverfahren (die SWD bietet lediglich das Synonym "Erwerbstätigkeit").



Abb. 7: Beispiel für einen Treffer zur Suchfrage "Frau und Beruf"

#### 4.3 Retrievaltest 3

Wie oben (3.2) erwähnt, wurden für den dritten Retrievaltest alle Suchformulierungen mit Oder-Verknüpfungen (43 von 100) in Suchfragen mit Verwendung *nur* der Singularform bzw. *nur* der Pluralform (letzteres in Einzelfällen) umgewandelt und diese neuerlich im Test-OPAC recherchiert. Alle Kennzahlen wurden neu berechnet, wobei sich aber zeigte, dass die modifizierten Suchanfragen kaum Änderungen im Gesamtergebnis bewirkten. Während die Suche im herkömmlichen Basic Index 1.032 Treffer erbrachte (zuvor 1.065), von denen 91,96% (91,8%) relevant waren, wurden bei der Suche im Basic Index mit Indexaten 1.361



(1.410) Dokumente mit einer Precision von 84,3% (83,1%) gefunden. Der Trefferzuwachs reduzierte sich sogar von 345 (= 32,4%) auf 329 (=31,9%). Dies zeigt, dass mit ODER verknüpfte Pluralformen bzw. (vereinzelt) Wortableitungen zumindest in dieser Untersuchung keine wesentlichen Effekte auf die Retrievalergebnisse hatten. Lediglich im Falle der Reduktion des Suchbegriffs "Wissenschaftliches Arbeiten" auf die Formulierung "wissenschaftlich AND arbeit" kam es zu einem deutlichen Trefferzuwachs, da mit solch einer realitätsfernen Suchformulierung zunächst überhaupt keine Treffer erzielt wurden. Selbst dieses (arbiträre) Resultat hätte das Gesamtergebnis nicht wesentlich beeinflusst.<sup>24</sup>

## **5 Resümee & Perspektiven**

### **5.1 Bewertung der Testergebnisse**

Der Test des automatischen Indexierungsverfahrens MILOS/IDX mit den Daten des Österreichischen Bibliothekenverbundes erbrachte im wesentlichen nur positive und erfolgversprechende Resultate:

- Der Trefferzuwachs erwies sich mit rund einem Drittel zwar nicht als sensationell, kann aber als durchaus respektabel bezeichnet werden.
- Von den durchschnittlich 3,45 aufgrund der automatischen Indexierung neu hinzugekommenen Treffern waren im Mittel (fast) zwei relevant, was in jedem Fall als Gewinn für die Benutzer bewertet werden kann.
- Die Zahl der Suchresultate ohne Treffer – oft ein besonders gravierendes Problem für OPAC-Benutzer – konnte durch die automatische Indexierung um immerhin ein Drittel gesenkt werden.
- Nach der automatischen Indexierung wurden erwartungsgemäss vermehrt auch Datensätze ohne RSWK-Ketten gefunden, die hinsichtlich ihrer Relevanz den Datensätzen mit einer verbalen Sacherschliessung kaum nachstanden.
- Dennoch erwies sich der Anteil der Sätze mit RSWK-Ketten auch bei der Suche im Basic Index mit den neuen Indexaten als überraschend hoch. Das Vorhandensein einer verbalen Sacherschliessung hat sich beim Einsatz des automatischen Indexierungsverfahrens als vorteilhaft erwiesen (eine automatische Indexierung kann daher nicht als vollständiger Ersatz für eine intellektuelle Erschliessung gelten).

- Der erzielte Zuwachs an relevanten Treffern bei gleichzeitig nur mässig reduzierter Precision und sinkender Nulltreffer-Zahl kann im grossen und ganzen als Bestätigung der früheren Studien über automatische Indexierung in deutschsprachigen OPACs gewertet werden.

## 5.2 Ausblick

Wie ein automatisches Indexierungsverfahren für die tägliche Arbeit in der Zentrale des Österreichischen Bibliothekenverbundes implementiert und in die sonstigen Betriebsabläufe eingebunden werden könnte, ist derzeit Gegenstand von Überlegungen. Dabei geht es etwa um die möglichen Szenarien für die automatische Indexierung des bereits vorliegenden Datenbestandes von rund 3,76 Millionen Titeldatensätzen (Oktober 2003) des Verbundkataloges – entweder in sukzessiven Teilschritten oder als Gesamtindexierung anlässlich eines Releasewechsels der Bibliotheksverwaltungssoftware – sowie um die Übernahme der automatisch generierten Indexate durch die am Verbund beteiligten Lokalsysteme. Des weiteren werden Modelle für periodische Updates der neu hinzukommenden und der modifizierten (durch die Katalogisierung erweiterten oder korrigierten) Datensätze unter Berücksichtigung der Verteilung an die Lokalsysteme diskutiert. Dies könnte z.B. jede Nacht automatisch auf Basis der Datumsfelder in den Datensätzen erfolgen, wobei der Zeitaufwand angesichts einer vermutlich überschaubaren Menge von Titeln (zur Zeit max. 5.000 pro Tag) als unkritisch betrachtet werden kann. Eine tägliche Verarbeitung würde in den meisten unserer Verbundbibliotheken auch dazu führen, dass zum Zeitpunkt der Durchführung der Sacherschliessung die Datensätze schon automatisch indexiert wären und somit den Sacherschliessern ein zumindest auf Basis der Titelwörter angereichertes Katalogisat zur Verfügung stünde ("computerunterstütztes Indexing"). Ausserdem könnten zu diesem Zeitpunkt allfällig vorliegende Fehler des Automatismus erkannt und korrigiert werden; dabei müsste ein noch zu spezifizierendes Markierungsverfahren verhindern, dass derartige Korrekturen wieder aufgehoben werden, sobald der betreffende Datensatz aufgrund des dadurch erfolgten Updates der Datumsfelder erneut das automatische Indexierungsverfahren durchläuft.<sup>25</sup>

Im Falle des Einsatzes der automatischen Indexierung im Echtbetrieb des Österreichischen Bibliothekenverbundes wird vorrangig über die zu verwendende Software zu entscheiden sein. Neben MILOS/IDX kämen z.B. auch CAI (AUTINDEX) oder weitere Produkte in Frage; eine "echte" Marktforschung dazu steht noch aus.<sup>26</sup> Die voraussichtlichen Kosten für die

Software sind in diesem Zusammenhang ebenso zu untersuchen wie auch sonstige Kostenfaktoren, die bei einem praktischen Einsatz zu erwarten sind (z.B. Hardware, laufender Betrieb).

Während sich die hier vorgestellte Untersuchung auf deutschsprachiges Titelmateriale beschränkte, enthält ein grosser Verbundkatalog natürlich auch Datensätze, die dem Kriterium "deutsch" nicht genügen (das sind solche, die weder einen deutschsprachigen Titel noch RSWK-Ketten aufweisen). Für diese fremdsprachigen Datensätze ohne Beschlagwortung, deren Anteil in unserem Verbund schätzungsweise rund 20% beträgt, gibt es nach derzeitigem Stand noch keine Lösung, was die automatische Indexierung anbelangt.<sup>27</sup>

Da es sich beim Österreichischen Bibliothekenverbund um einen homogenen Aleph-Verbund handelt, in welchem die Katalogisierung im zentralen System erfolgt und die Katalogisate erst dann auf alle nutzenden Lokalsysteme verteilt werden, stellt sich auch die Frage nach der Distribution der durch eine automatische Indexierung generierten Indexate auf die lokalen Systeme bzw. nach den Bedingungen für eine solche Verbunddienstleistung (Verfahrensregeln, Kostenbeteiligung usw.) Das Interesse der lokalen Bibliotheken an einer Belieferung mit solchen Indexaten muss erst erhoben werden. Dies inkludiert auch die Erörterung der Problematik, dass bei einer etwaigen Nichtnutzung der Indexate durch ein Lokalsystem die (Erwerbs-)Katalogisierung in dieser Bibliothek aufgrund unterschiedlicher Basic Indices beeinträchtigt würde.<sup>28</sup>

Abschliessend kann festgehalten werden, dass sämtliche hier dargestellten Testergebnisse für den Einsatz eines automatischen Indexierungsverfahrens zur Anreicherung des OPACs des Österreichischen Bibliothekenverbundes sprechen und daher alle weiteren Schritte zur Realisierung eines solchen Vorhabens sobald wie möglich getroffen werden sollten.

## **6 Danksagung**

Die Autoren sind insbesondere Herrn Prof. Dr. K. LEPSKY, Fachhochschule Köln, der die Anwendung von MILOS/IDX für die vorliegende Studie ermöglichte, zu Dank verpflichtet. Ausserdem danken wir unseren Wiener Kolleginnen und Kollegen, die am ersten Retrievaltest mitgewirkt, bei der Auswertung mitgeholfen und/oder die Untersuchung auf andere Weise unterstützt haben.

## 7 Anmerkungen

<sup>1</sup> Dass diese Systeme für Online-Kataloge nur bedingt geeignet sind, ist inzwischen in Fachkreisen hinlänglich bekannt und soll hier nur angemerkt werden.

<sup>2</sup> Einige vorliegende Zahlen: Verbunddatenbank Nordrhein-Westfalen: ca. 10% (vgl. Seiffert, F.: Aspekte zur Inhaltsermittlung im HBZ-Verbund. Vortrag, Köln, 04.12.2002 [siehe Anm. 13]; Südwestdeutscher Bibliotheksverbund: ca. 12% (2000), vgl. <http://www.bsz-bw.de/statistik/sach.html> [besucht: 27.10.2003]; Bibliotheksverbund Bayern: ca. 37% (2000), vgl. <http://www-opac.bib-bvb.de/subbvb/ordbvb/info/bestand.htm> [besucht: 27.10.2003].

<sup>3</sup> So z.B. der Österreichische Verbundkatalog (<http://opac.bibvb.ac.at/acc01>) und zahlreiche andere österreichische OPACs.

<sup>4</sup> Nohr, H. (2003). *Grundlagen der automatischen Indexierung: Ein Lehrbuch*. Berlin: Logos-Verl. (hier: S. 20).

<sup>5</sup> Im Gegensatz zu dokumentarischen Datenbanken, die oft auch Abstracts oder gar Volltexte – also ein viel umfangreicheres Wortmaterial – enthalten. Diese benötigen z.B. eine statistische Komponente zur Gewichtung der Suchbegriffe anhand von Termfrequenzen u.dgl.

<sup>6</sup> MILOS = **M**aschinelle **I**ndexierung zur verbesserten **L**iteraturschließung in **O**nline-**S**ystemen.

<sup>7</sup> Entwickelt von der Firma Softex (nunmehr Delta), basierend auf Arbeiten von Prof. HARALD ZIMMERMANN, (Fachrichtung Informationswissenschaft, Universität des Saarlandes).

<sup>8</sup> Lepsky, K. (1994). *Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen*. Köln: Greven. (Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen, H. 18). – Lepsky, K. (1996). Automatische Indexierung und bibliothekarische Inhaltsermittlung: Ergebnisse des DFG-Projekts MILOS I. In: Niggemann, E.; Lepsky, K. (Hrsg.) *Zukunft der Sacherschließung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995*. Düsseldorf: Universitäts- und Landesbibliothek. 12–36. – Lepsky, K.; Siepmann, J.; Zimmermann, A. (1996). Automatische Indexierung für Online-Kataloge: Ergebnisse eines Retrievaltests. *Zeitschrift für Bibliothekswesen und Bibliographie* 43(1). 47–56.

<sup>9</sup> Gödert, W.; Liebig, M. (1997). Maschinelle Indexierung auf dem Prüfstand: Ergebnisse eines Retrievaltests zum MILOS II Projekt. *Bibliotheksdienst*. 31(1). 59–68. – Sachse, E.; Liebig, M.; Gödert, W. (1998). *Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II-Projekt*. Köln: FH Köln. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft, Bd 14).

<sup>10</sup> Grummann, M. (2000). Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet? Retrievaltests von indexierten ekz-Daten mit der Software IDX. *Bibliothek: Forschung und Praxis*. 24(3). 297–318.

<sup>11</sup> Vgl. z.B. die Hinweise bei Schmidt, S. (2002). Verfahren zur maschinellen Indexierung von Dokumenten (MILOS). [Vorlesungstext, Kapitel 3; hier: Abschnitt 3.4]. Letzte Aktualisierung: 23.04.2002. Online: [http://www.dombibliothek-koeln.de/index1.html?publikation/schmidt/ss2002/kapitel\\_3.html](http://www.dombibliothek-koeln.de/index1.html?publikation/schmidt/ss2002/kapitel_3.html) [besucht: 07.10.2003].

<sup>12</sup> Vgl. dazu die Präsentationen bei der Arbeitstagung der Verbundbibliotheken im nordrhein-westfälischen Verbund, *Maschinelle Indexierung - innovative Verfahren zur Inhaltsermittlung im Verbund*, Köln, 04.12.2002. Letzte Aktualisierung: 30.09.2003. Online: <http://www.hbz-nrw.de/wir/publika/indexierung.html> [besucht: 17.10.2003].

<sup>13</sup> Lepsky, K.; Zimmermann, H.H. (2000). Katalogerweiterung durch Scanning und automatische Dokumenterschließung: Ergebnisse des DFG-Projekts KASCADE: Stand: 18.05.2000. *Zeitschrift für Bibliothekswesen und Bibliographie*. 47(4). 305–316. – Lohmann, H. (2000). *KASCADE: Dokumentanreicherung und automatische Inhaltsermittlung: Projektbericht und Ergebnisse des Retrievaltests*. Düsseldorf: Univ.- und Landesbibliothek. (Schriften der Universitäts- und Landesbibliothek Düsseldorf, 31).

<sup>14</sup> Hauer, M. (2003). Wissensressourcen zutage fördern: Digitalisierung von Aufsätzen und anderen Texten mit maschineller Inhaltsermittlung am Beispiel der Vorarlberger Landesbibliothek. *BuB Forum für Bibliothek und Information*. 55(3). 192–196. – "CAI-Engine" basiert auf dem von Prof. JOHANN HALLER et al.

---

am Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes entwickelten automatischen Indexierungsverfahren AUTINDEX.

<sup>15</sup> Darunter nahezu alle österreichischen Universitätsbibliotheken und die Österreichische Nationalbibliothek; vgl.: <http://www.bibvb.ac.at/verbund-opac.htm> [Webseite des Verbundes, laufend aktualisiert].

<sup>16</sup> Adresse des Verbund-OPACs: <http://opac.bibvb.ac.at/acc01>

<sup>17</sup> Dadurch führt bspw. die Eingabe des Nondesktors "Bibliotheksautomation" zu Treffern, die mit der SWD-Vorzugsbenennung "Bibliothek / Automation" beschlagwortet sind.

<sup>18</sup> Im Vorjahr hat z.B. eine Logfile-Analyse der UB Innsbruck erbracht, dass im dortigen OPAC rund drei Viertel aller Suchanfragen im Basic Index erfolgten; eine ähnliche oder gar noch höhere Quote kann für den Verbund-OPAC angenommen werden. Aus diesem Grund interessierte uns die Auswirkung auf den "Alle Felder"-Index mehr als jene auf spezifische Titel- oder Schlagwortindizes.

<sup>19</sup> Lepsky (1994, 32ff.); Nohr (2003, 66ff.); Zimmermann, H. H. (1996). Automatische Indexierung und elektronische Thesauri. In: Niggemann, E.; Lepsky, K. (Hrsg.) *Zukunft der Sacherschliessung im OPAC: Vorträge des 2. Düsseldorfer OPAC-Kolloquiums am 21. Juni 1995*. Düsseldorf: Universitäts- und Landesbibliothek. 37–47.

<sup>20</sup> Dabei hätte auch noch zwischen den beiden oben (2.3) erwähnten Kategorien differenziert werden können, doch wurde diese Möglichkeit in der vorliegenden Studie nicht verfolgt.

<sup>21</sup> Sachse, Liebig & Gödert (1998, 19–21).

<sup>22</sup> Sachse, Liebig & Gödert (1998, 31).

<sup>23</sup> Sachse, Liebig & Gödert (1998, 27–29).

<sup>24</sup> Der in der vorliegenden Studie im Vergleich zu MILOS II weniger spektakuläre Trefferzuwachs ist u.E. daher auf folgende Faktoren zurückzuführen: Bereits vorhandene "rudimentäre automatische Indexierung" (Synonyme aus der SWD); fehlende automatische Indexierung bei einem Teil der Testdaten aufgrund nicht korrekter automatischer Spracherkennung (der Anteil dieser Datensätze konnte leider nicht mit vertretbarem Aufwand quantifiziert werden, könnte aber in der Grössenordnung von einigen Prozent liegen); vermutlich unvollständige Ausnutzung des Leistungsumfangs von MILOS/IDX aufgrund der Projektrahmenbedingungen (vor allem im Bereich der Relationierungen, z.B. keine Oberbegriffe, keine Dekomposition der SWD-Verweisungsformen, keine Relationierung der als SWD-Verweisungen lexikalisierten Titelstichwörter); Suche im Basic Index anstatt in einzelnen Registern.

<sup>25</sup> Dies gilt auch für spätere Neuindexierungen, z.B. aufgrund verbesserter Wörterbücher oder Normdateien.

<sup>26</sup> Angesichts der Erfordernisse (deutsche Sprache, Eignung für OPAC-Daten, Einbindung der SWD) kann allerdings nicht erwartet werden, dass eine grosse Zahl von Produkten in Frage kommt.

<sup>27</sup> Die ursprünglich in IDX enthaltene wortbezogene Übersetzungskomponente wurde in den MILOS-Studien nach ersten Versuchen zurückgestellt und im EU-Projekt CANAL/LS weiterverfolgt (Lepsky, 1996, 17). – Lösungsmöglichkeiten sind eventuell in den Bestrebungen zur Internationalisierung der SWD zu sehen, z.B. im Projekt MACS (Multilingual Access to Subject Headings).

<sup>28</sup> Bei der unter Aleph 500 üblichen Katalogisierungspraxis wird zunächst eine Multipoolsuche (Lokaler Katalog, Verbundkatalog, Fremddatenbanken) durchgeführt, bei der z.B. ein im Verbundkatalog mittels automatischer Indexierung angereicherter Datensatz gefunden würde, ohne dass er im Ergebnisset für den nicht angereicherten Lokalkatalog aufschiene (Dublettengefahr!)